Are They Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere

Nuha Albadi[†*], Maram Kurdi[‡*], Shivakant Mishra[*]
[†]Department of Computer Science, Taibah University, Medina, Saudi Arabia
[‡]Department of Computer Science, Taif University, Taif, Saudi Arabia
[*]Department of Computer Science, University of Colorado Boulder, Boulder, USA
{nuha.albadi, maram.kurdi, mishras}@colorado.edu

***Abstract*—Religious hate speech in the Arabic Twittersphere is a notable problem that requires developing automated tools to detect messages that use inflammatory sectarian language to promote hatred and violence against people on the basis of religious affiliation. Distinguishing hate speech from other profane and vulgar language is quite a challenging task that requires deep linguistic analysis. The richness of the Arabic morphology and the limited available resources for the Arabic language make this task even more challenging. To the best of our knowledge, this paper is the first to address the problem of identifying speech promoting religious hatred in the Arabic Twitter. In this work, we describe how we created the first publicly available Arabic dataset annotated for the task of religious hate speech detection and the first Arabic lexicon consisting of terms commonly found in religious discussions along with scores representing their polarity and strength. We then developed various classification models using lexicon-based, n-gram-based, and deep-learning-based approaches. A detailed comparison of the performance of different models on a completely new unseen dataset is then presented. We find that a simple Recurrent Neural Network (RNN) architecture with Gated Recurrent Units (GRU) and pre-trained word embeddings can adequately detect religious hate speech with 0.84 Area Under the Receiver Operating Characteristic curve (AUROC).**

***Keywords*—cyberhate, religious hate speech, online radicalization, social media mining, text analytics, Arabic NLP, Twitter.**

I. Introduction

Twitter is one of the most widely used social networking sites in the Arab region with more than 11 million active users and over 27 million tweets a day as of March 2017 [1]. Unfortunately, Twitter and other social networking sites have been exploited by extremists who use derogatory and dehumanizing language to incite hatred and violence against religious groups around the globe [2]. In addition, hate speech on social media has been linked to an increase in physical hate crime incidents [3]. More worrying is that the rapid growth of social media has made it almost impossible to manually monitor and review the overwhelming number of daily messages posted online. Therefore, it has become particularly crucial to build tools that can automatically detect online hateful content without manual intervention to mitigate its harmful effects.

Although hate speech can be based on different protected characteristics, in this work our focus is on *religious* hate speech which we define as a speech that is insulting, offensive, or hurtful and is intended to incite hate, discrimination, or violence against an individual or a group of people on the basis of religious beliefs or lack of any religious beliefs. Interestingly, six of the eleven countries with the highest *Social Hostilities Index* [4], a measure of crimes partly motivated by religion, have Arabic as an official language. However, most prior work in the area of hate speech detection has targeted mainly English content [5], [6], [7], [8], [9], [10], [11]. Prior research in Arabic social media content has mostly focused on either detecting Jihadist/ISIS support messages [12], [13] or identifying vulgar/obscene language [14], which is distinguishable from hate speech. Figure 1 shows some typical examples of religious hate speech found in the Arabic Twitterspher.

Fig. 1. Examples of religious hate speech on Twitter

The complexity and richness of the Arabic morphology poses some unique challenges to Arabic NLP researchers [15]. In informal settings, as in social media, Dialectal Arabic is used more often than Modern Standard Arabic. Arabic has many different dialects varying not only from country to country but also from region to region within the same country. Dialectal Arabic, unlike Modern Standard Arabic, does not follow any standard grammar or spelling rules [15]. Similarly spelled words can have different meanings across different dialects, which increases the ambiguity of the language. For example, the word عافية *eafia* means "fire" in the Maghrebi Arabic, while it means "health" in the Gulf Arabic. More challenging is that Arabic is considerably an under-resourced language compared to English. One of these missing resources is the availability of an Arabic hate lexicon which can be very useful in cyberhate detection research.

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 ©

In addition to these unique challenges specific to Arabic, detecting religious hate speech is faced with the several challenges that are also encountered in detecting hate speech in English social media. The large volume of diverse content being posted on social media platforms makes it challenging to find common patterns and trends in data. Further user generated social network data contains noisy content such as incorrect grammar, misspelled words, Internet slangs, abbreviations, elongation of words, and text containing multi-lingual script, which poses technical challenges in text mining and linguistic analysis. Finally, social network guidelines typically prevent users from posting any illegal or unethical content. As a results, users post information which might seem genuine but leads to hate speech levels in a very subtle way. Again, this complicates building tools that can automatically detect religious hate speech.

In this paper, we investigate the problem of religious hate speech in Arabic Twittersphere and develop classifiers to automatically detect it. In particular, we collected 6,000 Arabic tweets referring to different religious groups and labeled them using crowdsourced workers. We provide a detailed analysis of the labeled dataset, reporting main targets of religious hatred in the Arabic Twitter space. After preprocessing the dataset, we applied various feature selection methods to create different lexicons consisting of terms found in tweets discussing religions along with scores reflecting their strength in distinguishing a sentiment polarity (hate or not hate). Finally, we report and compare the results of applying multiple classification approaches including lexicon-based, n-gram-based, and deep-learning-based methods on a new dataset to detect occurrences of religious hate speech.

The main contributions of this work are as follows:

1) To the best of our knowledge, this paper is the first research effort to tackle the problem of detecting religious hate speech on Arabic social media.
2) We create the first Arabic dataset annotated for the purpose of religious hate detection and the first Arabic lexicon of religious hate terms, making these resources public[1] to encourage further research in this domain.
3) We experiment with various classification models, and show that GRU-based deep neural network outperforms both n-gram-based and lexicon-based models.

II. Related work

There is some limited literature on the problem of misbehavior detection on Arabic social media. Magdy et al. [12] trained an SVM classifier to predict whether a user is more likely to be an ISIS supporter or opponent based on textual features of the user's tweets authored before declaring his/her support or opposition. Linguistic and temporal features have been used to detect Jihadist support instances on Twitter [13]. Mubarak et al. [14] proposed an approach for automatically creating and expanding a list of obscene words and then used the created list to detect profane tweets.

Hate speech has been investigated quite extensively in English social media content. Waseem and Hovy [8] suggested that character n-grams are better predictive features than word n-grams for recognizing racist and sexist tweets. In their study, they observed that by using gender as an additional feature resulted in minimal improvement to the classification results, while adding location information led to a decrease in performance. Their n-gram-based classification model was outperformed by a large margin using Gradient Boosted Decision Trees (GBDT) classifier trained on word embeddings learned using Long Short-Term Memory Network (LSTM) [7].

To identify the main targets of hate speech in social media, Silva et. al. [10] proposed to use sentences of the structure '*I <intensity> <intent> <group of> people*', where *<intent>* captures the word hate or one of its synonyms, and <group of> captures a single word that describes a particular group of people, e.g. Mexican. The problem of distinguishing hate speech from the general offensive language has been studied by Davidson et. al. [5], in which they showed that 31% of their hateful tweets were misclassified as offensive, while only 5% of their offensive tweets were mislabeled as hate.

III. Data

In this paper, we focus on the four most common religious beliefs in the Middle East. These include Islam (93.0%), Christianity (3.7%), Judaism (1.6%), and Atheism (0.6%) [16]. Since Islam is the most practiced religion in this region, we include the two main sects of Islam, namely Sunni and Shia which comprises 87-90%, and 10-13% of all Muslims respectively [17].

A. Data collection

In November 2017, using the Twitter's search API[2], we collected 6000 Arabic tweets, 1000 for each of the six religious groups. We used this collection of tweets as our training dataset. Due to the unavailability of a hate lexicon and to ensure unbiased data collection process, we included in the search query only impartial terms that refer to a religion name or the people practicing that religion. Specifically, we did not use any religious slurs that are used to insult people of a particular religious affiliation. For example, when collecting tweets related to Islam, we used the Arabic equivalent of the keywords: Islam, Muslim, and Muslims. To minimize redundancy in our training dataset, we collected only original tweets by excluding retweets from our queries. We also didn't collect any reply tweets to ensure that the tweets gleaned are self-contained to maximize the ability of crowdsourced workers to make reliable judgments.

In January 2018, we collected another set of 600 tweets, 100 for each of the six religious groups, for our testing dataset. We employed the same methodology for collecting this set as we used for collecting the training set. We intentionally collected a completely new unseen data, two months after we had collected our training data, to ensure reliable classification results and that the developed classifiers can generalize well to new data.

[1] https://github.com/nuhaalbadi/Arabic_hatespeech

[2] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

B. Annotation

We designed a task on CrowdFlower[3], a crowdsourcing platform, to obtain annotations for our training and testing datasets. We allowed only Arabic speaking annotators with IP addresses from one of the Arabic-speaking Middle Eastern countries to access the task. Before starting the annotation process, annotators were provided with our definition (stated in Section I) and some examples of religious hate speech. To reduce subjective biases, annotators were specifically asked not to allow their personal beliefs or religious affiliation influence their judgment.

We asked annotators to read a tweet carefully before deciding if the tweet: a) included an instance of religious hate speech (we refer to this as *hate* class); b) didn't contain any instances of religious hate speech (we refer to this as ¬*hate* class); c) was unclear or unrelated to religious hate speech. If annotators decided that the tweet contained an instance of hate speech, they were asked to select one or more religious groups that the tweet was being hateful to. The religious affiliations provided in the second question were as follows: Muslims, Jews, Christians, Atheists, Sunnis, Shia, and/or other.

To ensure high quality annotations, we created a set of 100 test questions to be used in the *Quiz Mode* and *Work Mode*. In the *Quiz Mode*, annotators were asked a set of four test questions; only those who scored an accuracy of 70% or more were able to qualify as annotators for our task. In the *Work Mode*, one test question was injected per page of work, and annotators who failed to maintain an accuracy of at least 70% throughout the task were disqualified and excluded from the task. Each page of work contained five tweets, one of which was a test question, and annotators didn't know which of the five tweets was the test question. For each tweet, we collected three trusted judgments from three different annotators. Untrusted judgments, from annotators whose accuracies have fallen down 70%, were excluded from our final results.

C. Analysis of ground truth data

Using this methodology, we obtained a total of 19,845 judgments from CrowdFlower workers, 304 of which were untrusted and consequently excluded. The trusted judgments were made by 234 different annotators. The first question had an average inter-annotator agreement of 81%, while the second question had an average inter-annotator agreement of 55%. This means that annotators usually agreed whether or not a tweet was a religious hate speech, but they disagreed sometimes in specifying which religious groups were targeted.

In analyzing answers to the first question, we selected the answer with the highest confidence, which is a score between 0 and 1 that reflects the level of agreement on a given answer among annotators. This score was weighted by annotators' trust scores, their accuracies on test questions. Results show that 42% of the tweets in our training dataset were regarded as religious *hate* speech, while 52% were considered ¬*hate*, and 6% were regarded as unclear/unrelated. Tweets that were labeled unclear/unrelated were considered noise and therefore removed from our dataset.

[3]https://www.crowdflower.com

This was quite a revelation that a very large part of discussion about religion in Arabic Twittersphere is about hatred towards religious groups. We are not aware of a similar measurement study in English Twitter space in which hate against different religions is quantified. However, a related study has been conducted by Magdy et al. [18] where they used crowdsourcing annotations and label propagation technique to annotate a little over 336,000 English tweets referring to Islam. These tweets were randomly sampled from a large collection of tweets responding to the November 2015 Paris terror attacks. They found that 22% of the tweets were attacking Islam, while 61% were actually defending Islam, and 22% were neutral.

For the second question, we considered only answers with a confidence score higher than 0.3. Given the few disagreements among annotators regarding which religious groups were being targeted, a higher confidence score would result in returning an empty selection for some of the tweets that are labeled as hate speech, i.e., increasing the confidence score resulted in having some tweets labeled as hate but had no targeted religious group. The relatively higher disagreements among annotators (compared to questions 1) could be attributed to the fact that the second question provided much room for disagreements as it had seven choices for annotators to select from, while question 1 had only three choices. An example of a case that resulted in disagreement is when a tweet was targeting Muslims, some annotators included Sunnis and Shia as they are considered Muslims as well, while others selected only Muslims as the tweet was targeting Muslims in general and not a specific Islamic denomination.

Figure 2 shows the percentage of tweets that were considered hate speech against a religious affiliation among the 1000 tweets collected for that religious affiliation. We find that Jews and Atheists have the highest percentages of tweets (60% for Jews and 56% for Atheists) labeled as *hate* toward them in their datasets. Shia ranks third with about half of the tweets mentioning Shia being considered religious hate speech against them. This shows that a very large part of discussion about Jews, Atheists or Shia in Arabic Twittersphere is about (or contains) hatred towards these groups. Although some of Christians' tweets were hateful toward them (36%), they were among the religious groups with the least percentages of hateful tweets, followed by Sunnis (12%) and Muslims (2%). The average of these percentages is less than 42% because some of the tweets collected for a particular religious affiliation were hateful but not toward that particular affiliation; such tweets were not considered in this analysis graph.

Next, we look at all the tweets that were assigned the class *hate*. There were 2,526 such tweets. We found that Jews were the main target of religious prejudice with 33% of all hateful tweets targeted against them (see Figure 3). Shia were the second most discriminated group given that 32% of all hateful tweets were regarded as hateful against Shia. Christians ranked third with 25% and Atheists ranked fourth with 24%. The least targeted groups were Muslims (9%) and Sunnis (7%). The sum of all percentages in this figure exceeds 100 because several tweets labeled as *hate* targeted more than one religion at the same time.

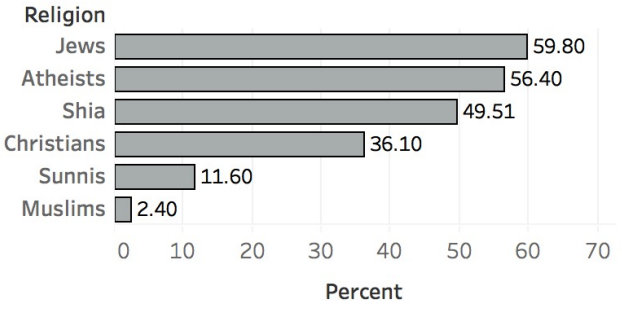

Fig. 2. Percentage of tweets labeled as *hate* against each of the religious groups when considering individually each of the 1000 tweets collected for each of the religious groups.

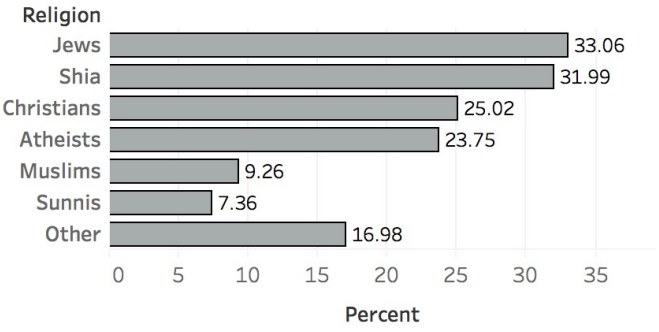

Fig. 3. Percentage of hateful tweets received by each of the religious groups among all hateful tweets.

D. Data preprocessing

We followed some of the Arabic-specific normalization steps proposed in [15] along with some other preprocessing steps that are necessary when working with microposts, tweets in particular. These preprocessing steps are as follows.

- Normalizing *Alef*: {أ, إ, آ, ٱ} → ا
- Normalizing *Alef Maqsoura*: ى → ي
- Normalizing *Ta Marbouta*: ة → ه
- Normalizing links, user mentions, and numbers to *somelink*, *someuser*, and *somenumber*, respectively.
- Normalizing hashtags by deleting underscores and the # symbol.
- Removing diacritics, *tatweel*, punctuations, emojis, non-Arabic characters, and one-letter words.
- Handling elongated words: It is common among social media users to repeat some letters to convey emphasis as in مشكووور *mashkuuuur* "thaaaanks". One approach for handling lengthening of words is to remove any repetition all together. However, such handling might remove consecutive letters from legitimate words that originally contain two consecutive occurrences of the same letter such as مملكة *mamlaka*. Although the method proposed in [15] for handling elongated Arabic words might give accurate results, the need for consulting a dictionary for every word makes it somewhat inefficient. Therefore, we only removed the repeated characters if the repetition was of count three or more. Also, we kept a record along each tweet of the number of elongated words that it contained, which can serve as a predictive feature for sentiment analysis.
- Stop words removal: We didn't remove any negation words since these are usually informative in sentiment analysis tasks [19]. The stop words were carefully selected so that we don't remove words that could change the meaning of a tweet. We cherry-picked some of the stop words provided by the ISRIStemmer [20], which is designed for only Modern Standard Arabic. To use it for our project, we supplemented the list to account for stop words used commonly in different Arabic dialects. For example, the Arabic word for "here" has at least five different dialectal variations, هنا *huna*, هني *hny*, هون *hwn*, هنيا *hania*, and هنايا *hanaya*. We also included stop words with common spelling mistakes found in social media posts, e.g. the word أنتِ *'int* is commonly misspell as أنتي *'anti*, so we included both in our list. In total, we have created a list of 356 stop words that we are sharing with the research community via our project GitHub repository[4].
- Stemming vs lemmatization: Both stemming and lemmatization are techniques for handling inflected words and reducing them to a common base form. Stemming chops off affixes based on predefined rules without the use of a dictionary, which can result in having a stem that is not a legitimate word in the dictionary. Lemmatization, on the other hand, tries to create a dictionary base form of the word (lemma) by using deep morphological analysis, a dictionary, and the context of the word in its reduction process. Thus, the meaning of a word is less likely to change when using a lemmatizer rather than a stemmer. We opt for lemmatization since one of our goals is to create a lexicon of religious hate terms. We use the state-of-the-art Arabic lemmatizer, MADAMIRA 2.1 [21] for lemmatization in our dataset.

IV. Lexicon generation

Sentiment lexicons have a variety of applications in the field of NLP such as sentiment analysis, information retrieval, and query expansion. In our training dataset, tweets are labeled as either *hate* or ¬*hate*. We Leveraged this labeled dataset to create three Arabic lexicons consisting of terms, each assigned a real-valued score reflecting their discriminative power toward a sentiment polarity. A positive score shows association with the *hate* class, while a negative score shows association with the ¬*hate* class. We experimented with three well-known feature selection methods to generate these lexicons:

- *AraHate-Chi* lexicon is generated using the chi-square (χ^2) [22] statistical test which measures the significance of association between a term and a class. At 95% significance level (p = 0.05) and degrees of freedom of one (df = 1), terms with scores whose

[4]https://github.com/nuhaalbadi/Arabic_hatespeech

absolute values are higher than 3.841 (P-Value ≤ 0.05) are statistically significant, i.e. they have a significant association with one of the classes. More details on how the χ^2 value is calculated can be found in [23] and [24].

- *AraHate-PMI* lexicon is generated using a scoring method based on Pointwise Mutual Information (PMI) [25] which measures the association strength between a term and a class. [23] explains in detail how PMI is used to build a domain specific lexicon.
- *AraHate-BNS* lexicon is created using Bi-Normal Separation (BNS) [26] sentiment scoring method which measures the predictive strength of a term in distinguishing a class label.

These feature selection methods have been specifically used due to their superiority and popularity in various classification tasks. In an empirical study [27], words selected by χ^2 resulted in achieving the best classification accuracy. PMI has been widely and successfully used in building domain-specific lexicons [28], [29], [30], [23], [24]. BNS has been shown to outperform other classic feature scoring metrics in a number of text classification tasks [31].

In all three approaches, we ignored words with frequency less than 10 as these might just be noise. In total, we have 1,523 words in each lexicon. Terms with the highest discriminative power toward *hate* and ¬*hate* classes using PMI, χ^2, and BNS are shown in Table I. We can see that BNS ranks words in a way similar to PMI but with different score values. Both PMI and BNS appear to be better than χ^2 in recognizing that some terms, e.g. religious affiliations, are not hateful words by themselves and therefore do not get assigned high scores. Words with negative scores, i.e. not hateful, are usually found in common Islamic prayers and supplications.

Upon analyzing terms with positive scores in all three lexicons, we can see that they generally fall into one of the following categories:

- Offensive/Vulgar: For example, وسخ "filthy", قذر "dirty", عميل "traitor", خائن "betrayer", خبيث "malevolent", عهر "prostitution", خنزير "pig", etc.
- Religious/Political: For example, كافر "infidel", مشرك "polytheist", خوارج "Khawarij", رافضه "Rafida", وهابي "Wahhabist", صهيوني "Zionist", ليبرالي "liberal", مجوس "Magi", صليبي "Crusader", etc.
- War/Violence: For example, عدو "enemy", دمر "destroy", قتل "killing", اباد "annihilated", اطرد "kick out", سقط "overthrow", عدوان "aggression" etc.

Terms in the Religious/Political category are often used pejoratively to describe people of certain religious affiliations. For example, the word وهابي "Wahhabist" is used pejoratively to refer to Sunni Muslims, while the word رافضه "Rafida" is used in a derogatory manner to refer to Shia Muslims.

As mentioned earlier, lexicons are typically used for sentiment analysis, information retrieval or query expansion. Our lexicons are based on the labeled dataset. To provide a quantitative evaluation of these lexicons, we test the discriminative power of lexicon in detecting religious hate speech. This is discussed in the next section.

V. Religious hate speech detection

In this section, we describe the different models we created to detect tweets with instances of religious hate speech. We used a separate unseen testing dataset for evaluating these models. For all models, both training and testing datasets have been preprocessed as described in Section III-D.

A. Experimental setup

The approaches we employed to detect religious hate speech can be categorized into three categories:

1) Lexicon-based approach: Each of the lexicons described in Section IV have been individually leveraged to detect hate speech by simply summing the sentiment scores (discriminative powers) of the tweet terms that exist in the lexicon; if a term appears in the tweet but does not exist in the lexicon, we assign that term a zero sentiment score. If the summation result is positive, we classify the tweet as *hate*, otherwise ¬*hate*. We used the result of this simple approach as a baseline for comparing and evaluating other classification models. We refer to the resulting classification models using the three various lexicons as *AraHate-PMI*, *AraHate-Chi*, and *AraHate-BNS*.

2) N-gram-based approach: We trained two classification models, namely logistic regression and SVM using n-gram model. We experimented with different n-gram features and different parameter settings, but we only report the best performing settings. The Logistic regression classifier was trained using character n-gram features (n = 1-4) with L2 regularization. The SVM classifier was also trained using character n-gram features (n = 1-4) with linear kernel and L2 regularization. We used Python sickit-learn library [32] to implement both models.

3) Deep neural network: Figure 4 illustrates the architecture of our GRU-based network with pre-trained word embeddings. First, we prepared the data by assigning integer indexes to unique words in our dataset. Tweets were then converted into sequences of integer indexes. These sequences were padded with zeros so that all sequences have an equal length of 50 (the longest tweet has 48 words). They were then fed into an embedding layer which maps word indexes to pre-trained word embeddings. We employed the *Twitter-CBOW* 300-dimension embedding model provided by AraVec [33] which contains over 331k word vectors that have been trained on about 67M Arabic tweets. The output of the embedding layer, an embedding vector of size (50, 300), was fed into a dropout layer with a rate of 0.5; the dropout layer was used as a form of regularization to prevent the model from overfitting. Then, a GRU layer with 240 hidden units was used to capture long-distance contextual information. The reason for using GRUs rather than LSTMs is that GRUs can train faster and may achieve a superior performance on datasets with limited number of training examples, i.e. GRUs may have better ability to generalize and less tendency to overfit small datasets [34]. The output layer was a 'sigmoid' layer that takes the output of the GRU layer, a vector of shape (1, 240), to predict the probability (from zero to one) of the tweet belonging to the positive (*hate*) class. We performed training

TABLE I
SNIPPET OF (A) *AraHate-PMI*, (B) *AraHate-Chi*, AND (C) *AraHate-BNS* LEXICONS SHOWING TERMS WITH HIGHEST PREDICTIVE POWER TOWARD *hate* CLASS (POSITIVE SCORES) AND ¬*hate* CLASS (NEGATIVE SCORES)

(a) *AraHate-PMI* lexicon

Term	Translation	score
laena لعنه	curse	+4.87
eahr عهر	whoredom/prostitution	+4.74
najas نجس	impure/filthy	+4.33
qarad قرد	monkey	+4.19
khinzir خنزير	pig	+4.19
libas لباس	garment	-4.52
akhlas اخلاص	sincerity	-4.60
'iibrahim ابراهيم	Abraham	-4.61
shifa' شفاء	healing	-4.71
eafia عافيه	health	-4.80

(b) *AraHate-Chi* lexicon

Term	Translation	score
yahudi يهودي	Jew	+274.65
mulahad ملحد	Atheist	+121.15
luein لعن	damn	+52.90
shiea شيعة	Shia	+52.32
'iirhab ارهاب	terrorism	+51.76
jmye جميع	all	-75.12
salam سلم	peace	-79.25
rahim رحم	mercy	-79.52
muslim مسلم	Muslim	-94.16
allahuma اللهم	O Allah	-151.68

(c) *AraHate-BNS* lexicon

Term	Translation	score
laena لعنه	curse	+1.88
eahr عهر	whoredom/prostitution	+1.84
najas نجس	impure/filthy	+1.71
qarad قرد	monkey	+1.66
khinzir خنزير	pig	+1.66
libas لباس	garment	-1.70
akhlas اخلاص	sincerity	-1.73
'iibrahim ابراهيم	Abraham	-1.74
shifa' شفاء	healing	-1.76
eafia عافيه	health	-1.79

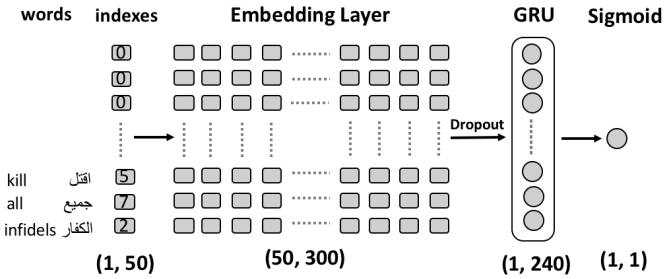

Fig. 4. Our GRU-based RNN architecture with pre-trained word embeddings

in batches of size 32, and we used 'adam' as our optimizer. It is worth mentioning that combining Convolutional Neural Network (CNN) with GRU didn't offer better results.

B. Results and discussion

The classification models were evaluated on a completely new unseen testing dataset, collected two months apart from the training dataset, as described in section III-A. Table II compares the performances of the various classification models in terms of F_1 score, precision, recall, accuracy, and Area Under the Receiver Operating Characteristic curve (AUROC). Highest scores are highlighted in **bold**.

From the table, we can see that the GRU-based RNN model achieved the best results with respect to all evaluation metrics. The 0.84 AUROC score achieved by the GRU-based model indicates that the model can separate the two classes reasonably well (see Figure 5). Among the lexicon-based methods, the *AraHate-PMI* provided the best results with respect to F_1 score, recall, and accuracy, while the *AraHate-BNS* performed better in terms of precision and AUCROC. The n-gram based models, logistic regression and SVM, showed similar performances to each other, but they generally performed better than the lexicon-based models, specially in terms of precision.

What makes the GRU-based RNN model perform quite better than all other models is its ability to develop deeper understanding of context and semantics. The use of a GRU layer with its gating mechanism allows for learning long-distance contextual relations that exist between words. Unlike n-gram models, word embeddings can capture the semantic relationships between words such as synonyms, antonyms, hyponyms, and hypernyms. Therefore, word embeddings allow semantically similar words to have similar dense vector representations. This in turn allows neural network models to gain deeper understanding of the natural language being processed.

Another advantage that the GRU model has over other models is that the word vectors that were used have been pre-trained on Twitter data where Dialectal Arabic and non-standard spellings is commonly used. This resulted in having 80% of the words in our dataset matched against the words in the word embedding model. To show that this 80% is considered a fairly good number, we report the results of using two other pre-trained word embeddings provided by Aravec [33]. The first word embedding model is *Web-CBOW* which has been learned on over 132M Arabic web pages where a mix of formal and informal Arabic is usually used. Using this model yielded lower match rate (71%) and lower performance (0.70 in F_1 score). The second word embedding model is *Wikipedia-CBOW* which was built on nearly 2M Arabic Wikipedia pages where mostly Modern Standard Arabic is used. Using this word embedding model resulted in even lower match rate (66%) and the worst performance of all word embeddings (0.66 in F_1 score).

Considering the 0.81 agreement score between annotators, we could say that the GRU-based model performed relatively well. One way to enhance the performance of the classifier would be to acquire more votes whenever there is a disagreement between annotators. Getting more training data may also help boost the performance of the model. Further analysis of the results shows that some of the misclassified cases were debatable ones, i.e., it was not very clear whether or not they should be regarded as religious hate speech. Other instances of religious hate speech that the classifier failed to detect were

TABLE II
EVALUATION RESULTS OF VARIOUS RELIGIOUS HATE SPEECH DETECTION MODELS

Model	F_1 score	Precision	Recall	Accuracy	AUROC
AraHate-PMI	0.69	0.66	0.72	0.71	0.78
AraHate-Chi	0.65	0.65	0.66	0.69	0.75
AraHate-BNS	0.64	0.67	0.62	0.70	0.79
logistic regression	0.71	0.71	0.70	0.74	0.81
SVM	0.72	0.72	0.72	0.75	0.81
GRU-based RNN	**0.77**	**0.76**	**0.78**	**0.79**	**0.84**

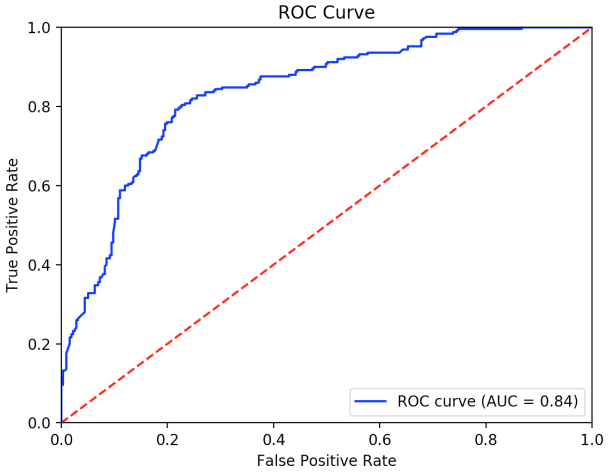

Fig. 5. The ROC curve for the GRU-based RNN model

those without explicit profane/offensive words.

Applying machine translation tools to leverage existing hate speech detection models trained on English content may not be efficient for several reasons. First, most existing Arabic-to-English machine translation tools target Modern Standard Arabic rather than Dialectal Arabic. Second, social media posts generally lack uniformity in writing styles and don't adhere to spelling/grammar standards; this nature of social media adds to the difficulty of developing reliable machine translation tools. Besides, hate speech is a subtle problem that is linguistically, culturally and historically dependent, which requires developing classifiers that capture these dependencies.

VI. CONCLUSIONS

In this paper, we provide a first attempt to investigate the problem of religious hate speech detection in Arabic Twitter. We have made several contributions to this problem. First, we created and published a dataset of 6,000 tweets labeled for this task. Second, we created and published three lexicons of religious hate terms, which can be used for various tasks, one of which is sampling microposts that may contain religious hate speech. Third, our analysis confirms that religious hate in Arabic Twitter space is very widespread. Nearly half of the discussions about religion in Arabic Twittersphere is about hate towards various religious groups, especially targeted towards Jews, Atheists and Shia. Finally, we investigated three different approaches to detect religious hate speech, namely lexicon-based, n-gram based, and deep learning-based approaches. The GRU-based RNN with pre-trained word embeddings gave the best performance with 0.79 accuracy and 0.84 AUROC.

There is still much room for improving the developed classifiers. We will carry out an extensive error analysis to identify challenges and limitations of each method. Although user features such as gender, age, education, job, etc., might not be readily available in social media, it can be predicted and investigated for correlation with religious hate speech. Another future direction is to investigate character-level neural networks since these are known to be effective specially when working with morphologically rich languages such as Arabic. We can further include other protected categories that are targeted in hate speech such as gender and race, and train a classifier that can distinguish between these protected characteristics.

REFERENCES

[1] F. Salem, "Social media and the internet of things towards data-driven policymaking in the arab world: Potential, limits and concerns," *Dubai: MBR School of Government*, vol. 7, 2017.

[2] Q. Wiktorowicz and S. Amanullah. (2015, February) How tech can fight extremism. CNN. Last accessed: March, 2018. [Online]. Available: https://www.cnn.com/2015/02/16/opinion/wiktorowicz-tech-fighting-extremism/index.html

[3] K. Müller and C. Schwarz, "Fanning the flames of hate: Social media and hate crime," 2017.

[4] Pew Research Center, Washington, D.C. (2017, April) Global restrictions on religion rise modestly in 2015, reversing downward trend - appendix b: Social hostilities index. Last accessed: March, 2018. [Online]. Available: http://assets.pewresearch.org/wp-content/uploads/sites/11/2017/04/07154135/Appendix-B.pdf

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ser. ICWSM '17, 2017, pp. 512–515.

[6] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proceedings of the 24th international conference on world wide web*. ACM, 2015, pp. 29–30.

[7] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 759–760.

[8] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.

[9] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 4, pp. 215–230, 2015.

[10] L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, "Analyzing the targets of hate in online social media." in *ICWSM*, 2016, pp. 687–690.

[11] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks." in *AAAI*, 2013.

[12] W. Magdy, K. Darwish, and I. Weber, "# failedrevolutions: Using twitter to study the antecedents of isis support," *First Monday*, vol. 21, no. 2, 2016.

[13] L. Kaati, E. Omer, N. Prucha, and A. Shrestha, "Detecting multipliers of jihadism on twitter," in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 954–960.

[14] H. Mubarak, K. Darwish, and W. Magdy, "Abusive language detection on arabic social media," in *Proceedings of the First Workshop on Abusive Language Online*, 2017, pp. 52–56.

[15] K. Darwish, W. Magdy, and A. Mourad, "Language processing for arabic microblog retrieval," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 2427–2430.

[16] Pew Research Center, Washington, D.C. (2015, April) Religious composition by country, 2010-2050. [Online]. Available: http://www.pewforum.org/2015/04/02/religious-projection-table/2010/percent/Middle_East-North_Africa/

[17] ——. (2009, October) Mapping the global muslim population. [Online]. Available: http://www.pewforum.org/2009/10/07/mapping-the-global-muslim-population/

[18] W. Magdy, K. Darwish, N. Abokhodair, A. Rahimi, and T. Baldwin, "# isisisnotislam or# deportallmuslims?: Predicting unspoken views," in *Proceedings of the 8th ACM Conference on Web Science*. ACM, 2016, pp. 95–106.

[19] R. M. Duwairi, R. Marji, N. Sha'ban, and S. Rushaidat, "Sentiment analysis in arabic tweets," in *Information and communication systems (icics), 2014 5th international conference on*. IEEE, 2014, pp. 1–6.

[20] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary," in *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, vol. 1. IEEE, 2005, pp. 152–157.

[21] A. Pasha, M. Al-Badrashiny, M. T. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth, "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic." in *LREC*, vol. 14, 2014, pp. 1094–1101.

[22] K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[23] N. Kaji and M. Kitsuregawa, "Building lexicon for sentiment analysis from massive collection of html documents," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[24] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "Crisislex: A lexicon for collecting and filtering microblogged communications in crises." in *ICWSM*, 2014.

[25] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.

[26] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1289–1305, 2003.

[27] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*. ACM, 2007, pp. 197–206.

[28] N. Al-Twairesh, H. Al-Khalifa, and A. AlSalman, "Arasenti: large-scale twitter-specific arabic sentiment lexicons," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 697–705.

[29] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Computational Intelligence*, vol. 31, no. 2, pp. 301–326, 2015.

[30] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets," in *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[31] G. Forman, "Bns feature scaling: an improved representation over tf-idf for svm text classification," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 263–270.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "Aravec: A set of arabic word embedding models for use in arabic nlp," *Procedia Computer Science*, vol. 117, pp. 256–265, 2017.

[34] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014. [Online]. Available: http://arxiv.org/abs/1412.3555